CRITICAL INFRASTRUCTURE PROTECTION: THREATS MINING AND ASSESSMENT

Giusj Digioia[(a)], Stefano Panzieri[(b)]

[(a) (b)] Roma Tre University – Via della Vasca Navale 79 – 00146 Rome Italy
[(a)] digioia@dia.uniroma3.it , [(b)] panzieri@dia.uniroma3.it

ABSTRACT
Within Homeland Defense, a crucial aspect is related to Critical Infrastructure (CI) protection. In fact, CIs encompass a wide range of strategic sectors for countries, such as food, water, public health, emergency services, energy, transportation, information technology, telecommunication and finance. Therefore, CIs operation should be granted to ensure national needs satisfaction. In order to monitor and prevent dangerous situations and threats that could affect CIs, Situation Awareness (SAW) theory addresses the goal of maintaining operator awareness, through rough data acquired by heterogeneous sensors monitoring CIs. One of the great problems in SAW is related to the definition of agile models able to highlight threats and to adapt themselves to monitoring requirements. This paper describes how Data Mining (DM) approach can be applied on data acquired by sensors watching at infrastructures, in order to build agile Hidden Markov Models (HMMs), for on-going situation assessment and consequent threat evaluation.

Keywords: critical infrastructure protection, data mining, situation awareness, agile hidden Markov models construction

1. INTRODUCTION

Within the context of homeland security, and in particular in critical infrastructure domain, one of the most felt issue is related to the availability of huge quantities of data acquired by sensors and the incapability to correlate relevant alarms, or to employ them for knowledge model construction that could lead to the definition of preventive actions and countermeasures.

Management of information stored in databases (DB) in order to discover hidden correlations, clusters of data and related descriptions is addressed by Data Mining (DM) discipline; while critical situation recognition and threat evaluation, starting from heterogeneous observations is an issue addressed by Situation Awareness (SAW) doctrine.

Both approaches deal with classification, but DM is generally applied off-line, on a set of data that can eventually be prepared for consequent statistical analysis. Such kind of data structuration leads to the definition of the so-called Data Warehouse (DW), where different classification techniques can be applied. Indeed, SAW manages on-line acquisition of data and try to fit them to a previously defined model, describing the domain of interest. In other words, DM allows extracting distinguishing features of a data; SAW tries to give a meaning to data, explaining why it has been gathered/observed.

The application of both approaches in the same system could be done as follows:

- adopting DM techniques to discover if and how information stored in DBs are correlated to a specific alarm (e.g. “power grid malfunctioning”);
- employing mined correlation to build knowledge models, related to most relevant alarms;
- adopting those models for Situation, Threat Assessment and Process Refinement (JDL levels 2, 3 and 4).

Researchers have started to study possible influences between Data Mining and Data Fusion domains only in recently published works.

In McConky [2012] data mining, and in particular an event co-reference process, is adopted to identify descriptions of the same event across sentences, documents, or structured databases. In McConky’s work the goal is to understand if different textual descriptions, stored in different DBs, refer to the same event, through the application of a customized event extraction technique and the evaluation of an event similarity measure. Also in the proposed architecture, the correlation between an event of interest and others, stored in different DBs, must be evaluated, but the focus is on causal, temporal and spatial correlation, rather than on similarities. Despite of McConky’s work, information managed in our work are properly structured for data mining classification, so that issues related to natural language interpretation are not taken into account. Moreover, while our paper presents an overview of a wide system architecture, in McConky [2012] the focus is on the implementation details of textual event description correlation and user Situation Awareness is supposed to be increased by presenting to him the collection of all existing descriptions of the same event. In this paper, SAW is regarded in a more complex view: user SAW is related to his capability to understand the undergoing situation and to evaluate its

threat. With this regard, the proposed system architecture supports the user through the application of inference algorithms on agile knowledge models, which are refined through relations mined in DB records.

In Stark [2012] it is proposed a mixed approach combining data mining and Bayesian Network approach, where BNs are built and validated employing data stored in DBs, and through a refinement process performed by the user. With this regard, Data Mining is meant as a learning process more than a way to discover implicit correlations among data, as instead it is regarded in this work.

In the work of Salerno [2004], from the comparison and analysis of JDL and Endsley's model for Situation Awareness, a new framework for SAW is proposed. Within this framework, Data Mining techniques are mentioned for their potentiality to discover relationships between entities in a database and employ them to generate predictive models capable of describing what has been examined in terms of an abstract mathematical formalism (usually, a graph-theoretic construct). Nevertheless, hints on Data Mining application within the architecture are not deepened.

Another case study in which data mining is applied to SAW is reported in Riveiro [2008], where data mining is integrated with information visualization techniques. The so called *visual data mining* approach aims to integrate the user in the knowledge discovery process using effective and efficient visualization techniques, in order to discover anomalies in maritime traffic.

Finally, Krishnaswamy [2005] presents an Advanced Driving Assistance System that analyses situational driver behavior and proposes real-time countermeasures to minimize fatalities/casualties. The system is based on Ubiquitous Data Mining (UDM) concepts. It fuses and analyses different types of information from crash data and physiological sensors to diagnose driving risks in real time. UDM is meant as the process of analyzing data emanating from distributed and heterogeneous sources, with mobile devices or within sensor networks.

This paper is organized as follows. Section 1 introduces this work, presents its motivations and related works, in Section **Errore. L'origine riferimento non è stata trovata.** an overview of the reference system architecture is described. Section **Errore. L'origine riferimento non è stata trovata.** presents how Data Mining and can be applied to Situation Awareness theory, and in particular for agile Hidden Markov Model definition. Finally Section 4 concludes this paper with remarks and future recommendations.

2. SYSTEM OVERVIEW

In this Section the overview of the reference system architecture is summarized. As mentioned before, the proposed architecture aims to combine Data Mining techniques to inference algorithms, specifically employed for Situation/Threat Assessment and pattern recognition.

The goal of Data Mining is to discover relations among data stored in different and huge databases, and to define clusters of records, characterized by similarities with regard to certain kind of relations. Situation Assessment methodologies allow to recognize situations of interest, observing data acquired real-time from different and heterogeneous sensors.

The link between the two approaches can be summarized as follows: Data Mining approach is employed to define correlations among data stored in databases and events/objects of interest for the user; mined correlations are employed to build knowledge models adopted in the Situation Assessment process.

Data Mining techniques employed in this work refer to *supervised classification*, where main features describing cluster of information are known and algorithm goal is to assess the belonging of data to each cluster. In particular, clusters will be defined according to temporal, spatial and causal relations.

Indeed, Situation Assessment techniques employed in this work refer to *Hidden Markov Model*, able to describe pattern of dynamic/time-dependent situations through a graph of nodes and edges, representing states and relations among them. In particular, the output of the data mining process (i.e. relations among data stored) is employed to build and refine iteratively the Hidden Markov Model adopted in the inference process, where observations from the field feed the model and allow to estimate the on-going situation, to project it and to evaluate its threat, according to the related impact. Finally, relations discovered in the Data Mining phase are regarded as cues for evidence search, contributing to JDL Level 4 functions, i.e. Process Refinement, Hall and Llinas [2001].

2.1. Architecture

In Figure 1 is depicted the overall system architecture. As it can be noticed, the inputs of the system are *observations gathered* from the field and a *Target Of Interest* (TOI), specified by the user. Observations are continuously stored in log databases and feed the Hidden Markov Model employed in the Situation Awareness process; the TOI represents a target of particular interest for a user, such as a specific alarm, a hardware component, or a particular event. When a user is interested in analyzing a specific TOI, to discover everything that could be correlated to it, that need to be under observation, and eventually to prevent related threats/malfunctioning, the user can define a TOI and give it as input to the proposed system.

The outputs of the process are basically a *TOI correlation tree* and a *Hidden Markov Model*. The correlation tree is built by the Data Mining Module that discovers different levels and kinds of correlations

among DB records and the specific TOI. The final TOI correlation tree is defined taking into account previous correlation trees built for the same kind of TOI.

HMM is generated by the Agile Model Construction Module, that translates relations discovered among DB records in the states and transitions of the HMM. The HMM is then employed for pattern recognition and threat assessment related to the specific TOI. The system contemplates also an Evidence Search Module, for SAW refinement.

When the system is operative, it manages a set of HMMs and related TOI correlation trees, corresponding to a set of different type of TOIs. Observations are employed to feed HMMs and alert the user about threats related to TOIs. Each time that a specific TOI becomes of greater interest (the user defines it as input of the system), a refinement process is started and leads to HMM and correlation tree refinement. In this way, models of the most critical TOIs are exactly those model most refined, accordingly to stored observations and previous analysis.

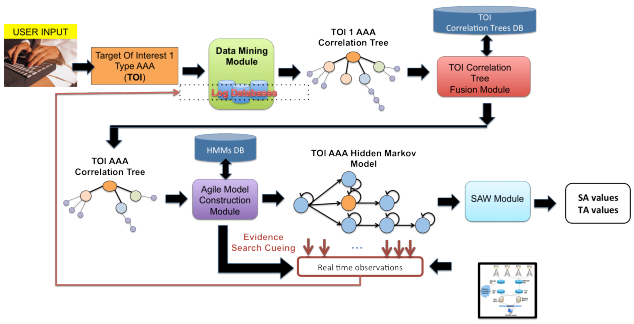

Figure 1 - System architecture

2.1.1. Architecture components

Main components of system architecture are listed below:

Target Of Interest (TOI) – It represents a target of particular interest for the user, as a particular alarm, a particular infrastructure component, a service, a particular event. For interest, we mean the need of the user to gather all correlated information from log DBs, to monitor it, and to prevent possible threats/malfunctioning related to it. The definition of TOIs depends on particular records stored in system DBs. A TOI can be characterized be the following information:

- *Type*: it specifies the typology of a TOI, such as event, component, service, alarm, etc. All TOIs of the same type refer to the same correlation tree and, consequently HMM. For example, "Deny of service" could be the type of the TOI related to a real cyber attack.
- *Name*: it describes the specific TOI.
- *Time*: this field define the temporal information related to the TOI, such ad the date, a time range, etc.
- *Location*: it indicates the geographical location related to the TOI.

Data Mining Module – The main purpose of this module is to mine correlations among records of log DBs and the TOI specified by the user. Data Mining techniques taken into account in this work refers to unsupervised and supervised classification, allowing to define data clusters, accordingly to a set of given variables. The variables chosen in this work express *temporal*, *causal* and *spatial* correlation.

- *Log Databases:* they represent the basis for the data warehouse the Data Mining Module is based on. They could be related to any kind of log, such as alarms, etc., but an intermediate process is required to prepare them to the following kind of analysis:
  - Temporal analysis*:* two records are temporally correlated if the distance between their temporal features is within a specific time range. In particular, different kind of temporal correlations can be defined: given a time interval among day, week, month, year, the records could be fully parallel, partially parallel, consequent.
  - Spatial analysis: two records are spatially correlated if the distance between their spatial feature is within a specific geographic range. In particular, different kind of spatial correlation can be defined: *city, region, country, or areas of different radius*.
  - Causal analysis: two records are causally correlated if the probability that A causes B is higher that a certain threshold δ: $P(B|A) = N_{B,A} / N_A > \delta$, where $N_{B,A}$ is the number of times B occurred within a certain time range, starting from the occurrence of A, and N_A is the total number of times A occurred.

TOI Correlation Tree – A correlation tree represents the tree of all records or TOIs correlated to a specific one, accordingly to at least one of the relations mentioned before. The root of a correlation tree is the TOI specified by the user, the nodes connected to the root with one link represent the 1st level correlated TOIs, the nodes that are distant 2 links from the root, represent the TOIs correlated with the first-level-correlated TOIs, and so on, see Figure 2. The weight of the links expresses the degree of correlation with the up-level node.

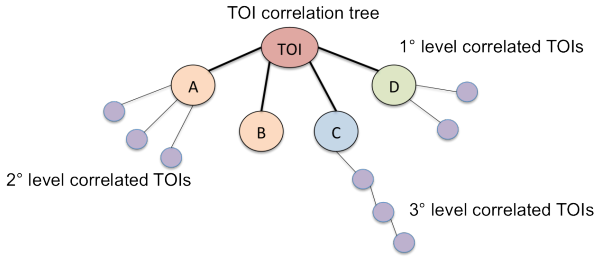

Figure 2 - TOI correlation tree

The size of the tree depends on the size of DBs and on the level of correlation to be investigated. TOI correlation trees can be related to specific TOIs, or to types of TOIs. In particular, the system first generates a specific TOI tree; then the Fusion Module compares it to the correlation tree of the TOI type; finally it updates the correlation tree for the specific type of TOI and stored it as a reference in a proper DB.

Agile Model Construction Module – this module goal is to derive, from the TOI type correlation trees, Hidden Markov Models with which monitor possible threats related to analyzed TOIs. HMMs of interest are therefore those related to TOI type like events or actions (i.e. "power grid malfunctioning"), whose threats need to be evaluated. Translation rules employed to build a HMM from a TOI correlation tree reflect the correlation type and weights mentioned above. In particular tree nodes correspond to the states of the HMM, while transition edges and probability reflects the weight of temporal, spatial and causal correlations among tree nodes. Details of this module are reported in next Section.

Evidence Cueing Module – Once that HMM is defined, a JDL level 4 refinement process can start, that is a process in which sensors and inference algorithms are refined in order to confirm/disconfirm the matching between observed reality and the knowledge model. In the proposed architecture, the refinement process is regarded as cueing the search for evidences that could be hidden to the observation process. In particular, the module highlights to the user the kind of evidence he should look for, accordingly to the type of TOI related to each node of the HMM.

SAW Module – The SAW module task is to take all observations coming from the field and to feed the HMMs stored in the HMM DB, in order to estimate on-going plans related to TOIs. As the approach adopted refers to Markov Theory, the inference process employs the Viterbi algorithm [James 2007], in order to estimate the most probable path, sequence of states, followed up to a certain time, given a sequence of observations. The probability of a certain path corresponds to the Situation Assessment value, expressing the confidence that a certain situation is undergoing. The SAW module then, computes the Threat Assessment value, expressing the impact that a certain situation could have: TA = SA value x damage caused.

In real-time context, two main processes can be identified, the SAW process and the Model Construction Process. The first one runs continuously, feeding HMMs stored in DBs with observations from the field, then archived in log DBs; the latter is triggered by the user, and in particular when he requires the construction and refinement of a specific TOI tree. The Model Construction Process affects the SAW process each time an HMM is updated. As noticed before, most refined models are related to TOIs of greater interest for the user.

3. DATA MINING FOR SITUATION AWARENESS

The process applied by the system at runtime can be summarized as follows:

1. DW update – the variables required for spatial, causal and temporal correlation with the TOI in input are computed, for all the DW records;
2. Unsupervised classification – cluster analysis is applied with regard to the variables added in the DW for spatial, temporal and causal correlation, in order to identify clusters in the DW.
3. Supervised classification – linear discriminant analysis is applied to define discriminant functions of all clusters identified in the unsupervised classification. Discriminant functions in the form of $g_i = \beta_1 x_{1i} + \ldots + \beta_p x_{pi}$ are computed for each record of the DW, and scores and coefficients are stored for TOI correlation tree construction.
4. TOI tree construction – for each cluster, a threshold is defined expressing the level of correlation of the record with the specific TOI. All records characterized by a score higher than the related threshold is added to at the first level of TOI correlation tree. The edge linking it with the TOI in input is characterized by:

    - The coefficients of the linear discriminant function that express how the record is correlated to the input TOI (for example, if the coefficient of the "Probability that the record is the cause of the specific TOI" is zero, it means that there is no causal correlation between the record and the TOI, while, if it is different from zero, it means that causal correlation exists and depends on the coefficient value).
    - The discriminant function score that expresses how much the record is correlated to the TOI, given the coefficient of the linear discriminant function, i.e. a specific kind of correlation (for example, causal and partially parallel temporal correlation).

The size of the TOI correlation tree can be increased repeating iteratively steps 1-4, using as input TOI a node of the just-generated correlation tree. Tree expansion could require high computation effort for the

system, therefore the expansion of the TOI correlation tree should be stopped at the very first levels, sufficient for the proposed system analysis.

A final consideration must be done on TOI correlation tree construction. The frequency with which TOIs are given in input to the system is supposed to be considerably lower than that with which observations are stored in the DW, and cluster analysis is strongly influenced by data employed by supervised and unsupervised classification. This means that new observations could lead to changes in cluster models that are effectively applied only when a specific TOI is given in input to the system. As HMMs for SAW are refined only when a TOI correlation tree is updated, independently by the observation storage process, a HMM will be refined as much as the related TOI is of interest for the user.

3.1. Unsupervised and supervised classification

Data Mining (DM) goal is to select, analyze and model huge quantities of data, in order to discover underlying and hidden relations of specific interest for the DB owner. DM can be regarded as an inductive methodology for information/knowledge retrieval from empirical data to general and theoretic rules to apply in wider contexts, in order to achieve a certain goal, rather than simple knowledge modeling, Nisbet et al. [2009].

In the proposed system, both unsupervised and supervised classifications are applied. Unsupervised classification aims to identify clusters of data accordingly to relevant variables. Usually, in order to identify the most discriminant components, the Principal Component Analysis (PCA) is applied. PCA computes eigenvectors of the correlation matrix and identify those components that classify the best the orthogonal projection of data analyzed. In the proposed system, the variables employed for unsupervised classification are already defined as those computed for causal, temporal and spatial comparison with the TOI specified by the user.

Once clusters of data are identified, through supervised classification, the linear discriminant analysis is applied in order to compute for each record a function of the form: $g_i = \beta_1 x_{1i} + \ldots + \beta_p x_{pi}$, where g represents the correlation score, and the coefficients are the components of the eigenvector, related to the maximum eigenvalue of the following matrix $D_W^{-1}(D_B)$, where D_W is the matrix expressing the deviation within a cluster, wile D_B is the matrix expressing the deviation between clusters. In order to get a good classification, the expression *Deviation Between/Deviation Within* should be maximized (considering the maximum eigenvalue), exactly how the linear discriminant analysis does. In this way, records of the same clusters are as much as possible similar among them and dissimilar with regard to other records of different clusters.

3.2. Situation Awareness

Situation Assessment is the process whose aim is situations recognition, by employing sequences of observations from the field. Observations usually are gathered from heterogeneous sensors, they are not always synchronized, and they can carry uncertain information. Robust SA techniques can efficiently manage such a kind of observation sequences, in order to fit them into a specific model of the observed reality.

A felt issue in the SA domain is related to model definition and refinement. In fact, if the knowledge model employed does not match with the domain of interest, SA algorithms would never provide relevant results.

In this work, the SA technique adopted refers to *Markov Theory* and contemplates the refinement of the knowledge model on the base of hidden relations existing on data and discovered by the Data Mining Module.

The employment of Hidden Markov Models (HMM) for pattern recognition is motivated by HMM capability to model plans, causal, temporal relations among states and by the existence of well known algorithms for path estimation among model states.

3.3. From Data Mining to HMMs

A crucial aspect of this work is related to the definition of HMMs employed by the system for SA. The task is absolved by the Agile Model Construction Module that is supposed to take in input the TOI correlation tree of a specific type, to look into the DB for the related HMM, and to substitute or to newly generate the corresponding HMM to be employed in SA.

Note that HMMs are generated only for those TOIs representing dangerous events or actions that could be prevented, if adequately monitored. Each time the user considers as relevant a specific event, related HMMs are refined so that they can adapt themselves to new knowledge learnt. The refinement is not strictly limited to parameter tuning, James, Z. [2007], but involves the whole HMM structure, and this is what makes the mentioned HMMs agile.

In the construction of the model, the TOI in input represents the end node of the HMM; to each node in the correlation tree corresponds a state in the HMM; while to each link in the correlation tree corresponds a set of transition edges, accordingly to the cluster correlation model. If a node in the tree is correlated to the TOI in input, this means that it belongs to a specific cluster and that its correlation value is higher than a certain threshold. The model of the cluster is summarized by the coefficients of the linear discriminant function, estimated by the supervised classification. The coefficients are then used to weight and define edges in the HMM.

4. CONCLUSIONS

In this paper a system architecture combining Data Mining with Situation Awareness approaches has been presented. Data mining techniques are adopted to mine hidden relations among data stored in databases. The output of the Data Mining process is then used to build knowledge models employed for situation and threat assessment. Both databases and Situation and Threat Assessment processes are fed with observations gathered from the field.

Data Mining process described in this paper is triggered by the specification of a particular target of interest (TOI) by the user. Once a TOI is given as input, the system Data Warehouse (DW) is updated through the computation of variables expressing spatial, causal and temporal correlation between each record and the TOI itself. Then the unsupervised classification is performed in order to identify clusters of data that are characterized by the same kind of correlation with the specific TOI. When clusters have been identified, the supervised classification, and in particular the linear discriminant analysis, is applied to estimate models of each cluster as linear functions whose coefficients expresses the dependency of the correlation score with the spatial, causal and temporal variables, introduced in the DW.

The output of the Data Mining process (clusters of records and their models) is employed to build Hidden Markov Models for the recognition of situations of interest. The proposed architecture contemplates the refinement of knowledge models each time user requires the analysis of a certain TOI. Therefore, HMMs related to most critical TOIs are also those more frequently refined.

Further works will be addressed to the implementation and validation of the proposed methodology.

REFERENCES

McConky, K., Nagi, R., Sudit, M., Hughes, W., 2012. Improving Event Co-reference By Context Extraction and Dynamic Feature Weighting. *Proc. IEEE Multidisciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support*, 38-43.

Stark, R.F., Farry, M., Pfautz, J., 2012. Mixed-Initiative Data Mining with Bayesian Networs. *Proc. IEEE Multidisciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support*, 107-110.

Salerno, J., Hinman, M., Boulware, D., 2004. Building a Framework for Situation Awareness. *Proc. IEEE Information Fusion*, 107-110.

Riveiro, M., Falkman, G., Ziemke, T., 2008. Improving maritime anomaly detection and situation through interactive visualization. *Proc. IEEE Information Fusion*, 46-54.

Krishnaswamy, S., Loke, S.W., Rakotonirainy, A., Horovitz, O., and Gaber, M.M., 2005. Towards Situation Awareness and Ubiquitous Data Mining for Road Safety: Rationale and Architecture for a Compelling Application. Proc. of Conference on Intelligent Vehicles and Road Infrastructure, 16-17.

Hall, D.L. and Llinas, J., 2001. Handbook of multisensor data fusion, *CRC Press*.

James, Z., 2007. *Hidden Markov Model with Multiple Observation Processes*. Honour Thesis.

Nisbet, R., Elder, J., Miner, G., 2009. *Handbook of Statistical Analysis & Data Mining Applications*, Academic Press, Elsevier.